Conserved queen pheromones in bumblebees: a reply to Amsalem et al.

Luke Holman[1], Jelle S. van Zweden[2], Ricardo C. Oliveira[2], Annette van Oystaeyen[3] and Tom Wenseleers[2]

[1] School of Biosciences, University of Melbourne, Melbourne, Victoria, Australia
[2] Laboratory of Socioecology and Social Evolution, Zoological Institute, University of Leuven, Leuven, Belgium
[3] Research and Development, Biobest Belgium NV, Westerlo, Belgium

ABSTRACT

In a recent study, *Amsalem, Orlova & Grozinger (2015)* performed experiments with *Bombus impatiens* bumblebees to test the hypothesis that saturated cuticular hydrocarbons are evolutionarily conserved signals used to regulate reproductive division of labor in many Hymenopteran social insects. They concluded that the cuticular hydrocarbon pentacosane (C_{25}), previously identified as a queen pheromone in a congeneric bumblebee, does not affect worker reproduction in *B. impatiens*. Here we discuss some shortcomings of Amsalem et al.'s study that make its conclusions unreliable. In particular, several confounding effects may have affected the results of both experimental manipulations in the study. Additionally, the study's low sample sizes (mean n per treatment = 13.6, range: 4–23) give it low power, not 96–99% power as claimed, such that its conclusions may be false negatives. Inappropriate statistical tests were also used, and our reanalysis found that C_{25} substantially reduced and delayed worker egg laying in *B. impatiens*. We review the evidence that cuticular hydrocarbons act as queen pheromones, and offer some recommendations for future queen pheromone experiments.

Subjects Animal Behavior, Evolutionary Studies
Keywords Eusociality, Cuticular hydrocarbons, Fertility signals, Reproductive division of labour

Submitted 17 May 2016
Accepted 18 April 2017
Published 16 May 2017

Corresponding author
Luke Holman,
luke.holman@unimelb.edu.au

Academic editor
Dezene Huber

Additional Information and Declarations can be found on page 10

DOI **10.7717/peerj.3332**

INTRODUCTION

Over the last 20 years, evidence has accumulated that specific cuticular hydrocarbons (CHCs), which consistently differ between fertile and non-fertile colony members, help to regulate reproductive division of labour in eusocial ants, bees and wasps (e.g., *Monnin, Malosse & Peeters, 1998*; *Liebig et al., 2000*; *Dietemann et al., 2003*). This theory originally rested on indirect evidence, including observations that queens and workers apparently always differ in their CHC profiles (reviewed in *Van Oystaeyen et al., 2014*), that CHC profiles correlate with inter-individual variation in fecundity within a given caste (e.g., *D'Ettorre et al., 2004*; *Holman, Dreier & D'Ettorre, 2010*), and that workers can discriminate between the CHCs of fertile and non-fertile individuals (*Dietemann et al., 2003*; *D'Ettorre et al., 2004*). Recently, studies using synthetic hydrocarbons have experimentally demonstrated that queen-like CHCs affect worker ovarian development (in seven species; *Van Oystaeyen et al., 2014*; *Holman et al., 2010*; *Holman, Lanfear & D'Ettorre, 2013*; *Holman, Hanley & Millar, 2016*; *Holman, 2014*; *De Narbonne et al., 2016*;

Oi et al., 2016), and/or induce behavioural changes in workers that are putatively related to reproduction (in three species; *Holman et al., 2010*; *De Narbonne et al., 2016*; *Smith, Millar & Suarez, 2015*; *Smith, Hölldobler & Liebig, 2009*). A recent comparative analysis of chemicals thought to be correlated with caste or fertility in 64 species of social Hymenoptera concluded that these chemicals were most commonly saturated CHCs, and that the correlation between saturated CHCs and female fecundity appears to be ancestral in Hymenoptera (*Van Oystaeyen et al., 2014*). Because eusociality evolved several times in this clade, this result suggests that queen pheromones evolved from chemical signals or cues that were already present in the solitary common ancestor of bees, ants and wasps, which lived *c.* 145 million years ago (*Van Oystaeyen et al., 2014*).

Recently, *Amsalem, Orlova & Grozinger (2015)* performed bioassays with synthetic hydrocarbons in *Bombus impatiens* bumblebees to test the hypothesis that saturated CHCs are evolutionarily conserved signals used to regulate reproductive division of labour. In the congeneric bumblebee *B. terrestris*, two earlier experiments concluded that workers resorbed oocytes more often (*Van Oystaeyen et al., 2014*) and had fewer developing oocytes in their ovaries (*Holman, 2014*) after treatment with the queen-characteristic cuticular hydrocarbon pentacosane (C_{25}), leading those studies to conclude that C_{25} was a queen pheromone. Amsalem et al. reported that worker reproduction was unaffected by C_{25}, and also not affected by two other cuticular hydrocarbons referred to as "controls" (C_{23} and C_{27}—though these CHCs also correlated with fecundity, and so should perhaps instead be regarded as putative queen pheromones). Because the three hydrocarbons had no statistically significant effect on worker reproduction, Amsalem et al. concluded that saturated hydrocarbons associated with fertility do not affect worker reproduction in *B. impatiens*, and suggested that the theory presented above be reconsidered.

Although we are excited to see new experimental data in this area, we feel that Amsalem et al.'s conclusions are not justified by their data. We first point out some methodological problems with the study, then present a statistical reanalysis of its data. We conclude that the study does not conclusively demonstrate that the three fertility-linked hydrocarbons C_{23}, C_{25} and C_{27} are not pheromones, as it claimed. Although confounding effects make the data difficult to interpret, the new results tentatively suggest that worker fecundity is reduced following exposure to queen CHCs. We conclude with some suggestions for the design of future experiments.

METHODOLOGICAL ISSUES

We believe there are three methodological shortcomings in the new study. Firstly, Amsalem et al. aimed to test whether workers' responses to queen pheromones involve learning. They did this by examining responses to synthetic hydrocarbons in both "experienced" and "naïve" workers, so-called because the experienced workers had spent more time in a nest containing a live queen. However, the experienced/naïve treatment was confounded, and we believe that it provides only limited information about the role of learning in the response to queen CHCs. Amsalem et al. did not take the conventional experimental approach of starting with a common pool of individuals and then randomly dividing them

Table 1 Sample sizes in Amsalem et al.'s experiment. The table highlights that sample sizes were low and uneven, that certain colonies are over-represented in particular hydrocarbon treatments, and that the naïve and experienced treatments used mixed-colony or single-colony groups of workers, respectively. Note that we give the sample size in terms of the number of colony fragments, which is appropriate for the colony-level variables 'egg number' and 'latency to egg laying'. For response variables measured at the level of individual workers (i.e., presence of 'ready-to-lay' eggs, length of terminal oocycte, and oocyte resorption) the sample sizes are c. 3-fold higher, because each colony fragment contained three workers.

		Colony								Total *n*
		a	b	c	d	e	f	g	Mix of 1–3 colonies	
Control	Experienced	8	2	–	3	1	3	6	–	23
C_{23} High	Experienced	8	–	–	2	4	2	4	–	20
C_{23} Low	Experienced	–	–	–	–	5	1	4	–	10
C_{25} High	Experienced	10	–	–	4	4	2	3	–	23
C_{25} Low	Experienced	8	1	1	1	4	1	5	–	21
C_{27} High	Experienced	7	–	–	3	4	2	4	–	20
C_{27} Low	Experienced	–	–	–	–	4	2	4	–	10
Average		8.2	1.5	1	2.6	3.7	1.9	4.3		18.1
Control	Naïve	–	–	–	–	–	–	–	16	16
C_{23} High	Naïve	–	–	–	–	–	–	–	6	6
C_{23} Low	Naïve	–	–	–	–	–	–	–	4	4
C_{25} High	Naïve	–	–	–	–	–	–	–	13	13
C_{25} Low	Naïve	–	–	–	–	–	–	–	12	12
C_{27} High	Naïve	–	–	–	–	–	–	–	6	6
C_{27} Low	Naïve	–	–	–	–	–	–	–	6	6
Average									9.0	9.0

between the two learning treatments, but rather used two different sets of workers to set up the two treatments. As a result, the naïve workers were younger and larger, which likely explains why they had larger oocytes and a longer latency to egg laying irrespective of CHC treatment. This means that any effect of the naïve/experienced treatment on the responsiveness to queen CHCs could be due to differences in age, size or reproductive physiology rather than learning. Moreover, the colony fragments of experienced workers contained three workers from the same colony, while the naïve colony fragments contained workers from an unspecified mixture of one, two or three different colonies. The effect of exposure to foreign vs same-colony workers on the response to queen pheromone is untested, so this might be a problem. Finally, the sample size in the naïve treatment was twofold lower than in the experienced treatment (Table 1). This means that the differences in p-values between experienced and naïve bees may reflect a difference in statistical power (as reflected in the differing widths of the confidence intervals in our Fig. 1), rather than lower responsiveness in the naïve bees, as was claimed.

Secondly, the allocation of workers to CHC treatments was imbalanced such that different hydrocarbon treatments contained workers from different colonies (Table 1). This unequal allocation could have biased the results, because colonies differed in body size (Table S1A), which in turn produced significant differences in body size between the seven hydrocarbon treatments (Table S1B). Because body size correlated with most of

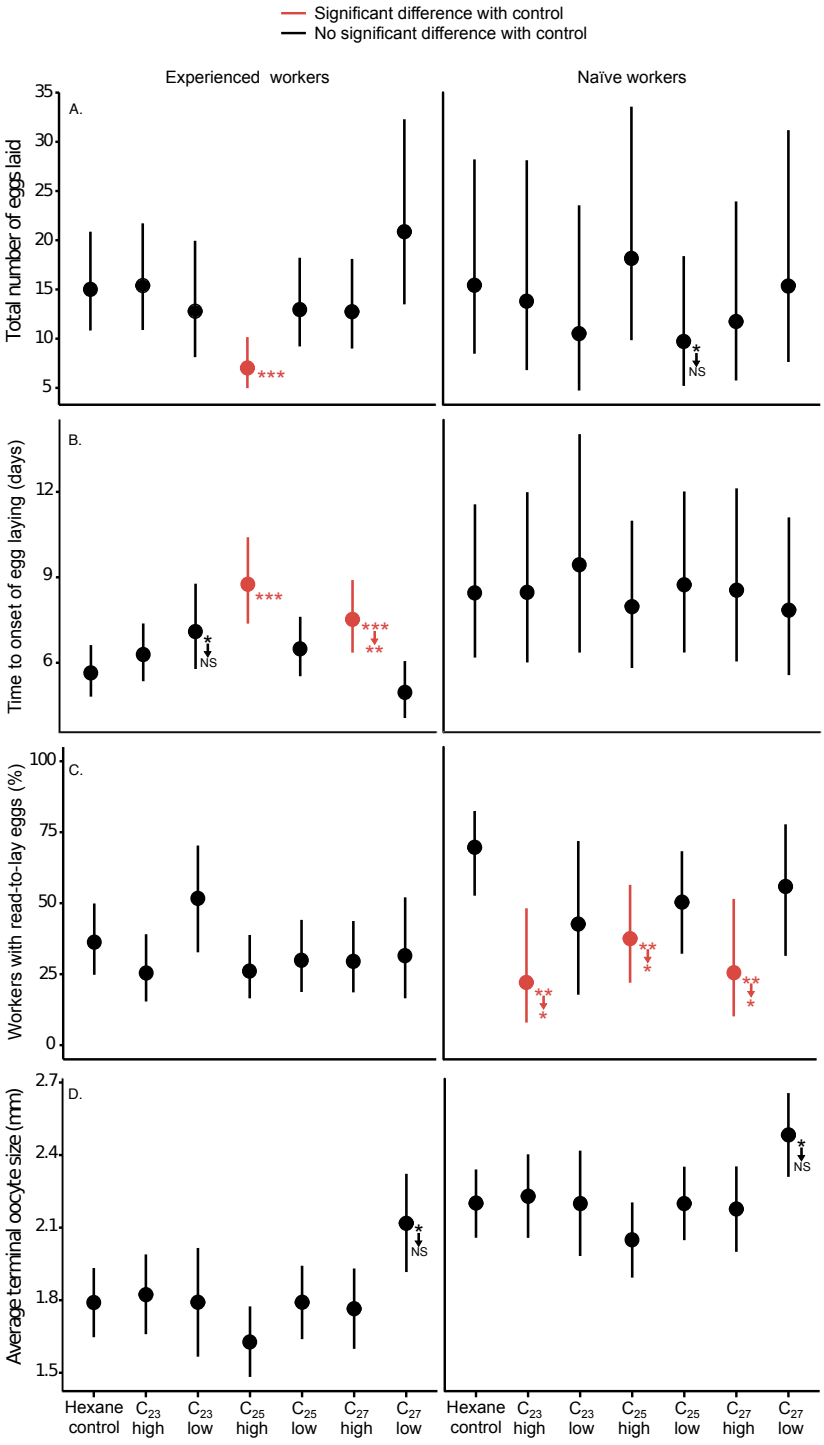

Figure 1 The least-square means (and their 95% confidence limits) for each hydrocarbon treatment, calculated from models shown in Tables S4–S7. Significant 2-tailed differences with the appropriate hexane control in planned contrasts are indicated using asterisks (***, $p < 0.001$, **, $p < 0.01$, *, $p < 0.05$; NS, $p > 0.05$). Results for which the significance level changed following Benjamini–Hochberg false discovery rate correction are indicated with downward arrows pointing to the new significance level. The results of corresponding fixed effect GLMs and Bayesian GLMMs were largely concordant and are presented in Tables S4–S7.

the variables under study (Tables S4–S8), the allocation of different-sized workers from various colonies to particular treatments may have skewed the results.

Thirdly, the 'Low' pheromone dose used in *Amsalem, Orlova & Grozinger (2015)* is very low relative to all previous queen pheromone experiments (see Table S2 for a review), and might be artificially low relative to natural conditions. Amsalem et al. chose not measure the mass of CHCs produced by queen or worker *B. impatiens*, and so it is unclear whether or not their experiment used a natural dose. Therefore, we analysed the cuticular hydrocarbon profiles of *B. impatiens* queens and workers using GC-MS (see Supplementary Methods), to obtain this information. We found that the 'Low' doses in Amsalem et al. represented approximately one ten-thousandth of the mass of CHCs found on the cuticle of a live queen (Table S2). In terms of ''queen equivalents'', this is approximately 140× lower than in *Holman (2014)* and 33,000× lower than the doses used in *Van Oystaeyen et al. (2014)* (Table S2), which may explain any difference in results between Amsalem et al. and the studies it set out to challenge. Although we acknowledge that it is difficult (if not impossible) to identify a biologically meaningful dose of queen pheromone in this type of experiment, we propose that the unusually low CHC concentration (especially in the ''Low'' treatment) may explain some of the new paper's null results.

STATISTICAL ISSUES

Firstly, *Amsalem, Orlova & Grozinger (2015)* applied classical ANOVA to types of data that can violate ANOVA's assumptions, including censored time series data and count data, which do not match the expected theoretical distributions and do not allow for appropriate bounds on the measurement scale. To address this, we instead analysed the data using generalized linear or generalized linear mixed models (GLM and GLMM) that make distributional assumptions matching those that are theoretically expected (e.g., Poisson models for count data) and which respect the bounds of the measurement scale by using an appropriate link function (e.g., the log link used in Poisson models ensures that counts are strictly positive). For time series data, we used survival models that appropriately model censoring in the data.

Secondly, Amsalem et al. searched for a significant difference between all possible pairwise combinations of the seven CHC treatment groups for multiple different metrics of worker reproduction, which results in a very large number of post-hoc tests. Because the authors corrected for multiple testing, the excessive number of tests greatly reduced statistical power. To address this problem, we used planned contrasts to compare each of the 6 CHC treatments to their respective control (i.e., 12 tests per ovary metric, because of the experienced/naïve treatment). Comparisons between a CHC and the control group are most informative for testing whether that CHC induces worker sterility—it is less interesting to test, for example, whether the low dose of C_{23} had a different effect to the high dose of C_{27}.

Thirdly, most of Amsalem et al.'s analyses do not statistically account for the non-independence introduced by the use of workers derived from the same colonies (or colony fragments, for the individual-level response variables), meaning that their analyses incur

pseudoreplication. One exception appears in Table S7 of *Amsalem, Orlova & Grozinger (2015)*, in which the authors fit ANOVAs with CHC treatment, colony, and their interaction as fixed effects. Because of the imbalanced experimental design, the authors chose to discard all the data from colonies a-d (i.e., 46% of the dataset; Table 1) in order to fit the treatment × colony interaction. If one instead uses all the data (which necessitates omitting the interaction term), one recovers the significant CHC treatment effects on worker egg laying, latency to egg-laying and worker ovary activation that we report below.

Fourthly, as mentioned above, worker body size differed substantially across pheromone treatments (Tables S1A and S1B), likely because of non-random treatment allocation, which is problematic because fecundity correlates with body size (Tables S4–S8). We therefore decided to statistically correct for this confounding difference in body size by including it as a covariate in our analyses (however, omitting body size did not qualitatively change the results we present here). Our models thus estimate the effect of treatment on the body size-corrected response variables.

Amsalem et al. reported on four response variables pertinent to the hypothesis that queen hydrocarbons affect worker reproduction: number of eggs laid (over 10 days), days until the onset of egg laying (censored after day 10), frequency of oocyte resorption, and mean size of the terminal oocytes. To facilitate comparison with past queen pheromone experiments, many of which treat ovary activation as a binary variable, we also coded a fifth response variable that was discussed in *Amsalem, Orlova & Grozinger (2015)* but not formally analysed: the frequency of workers with fully activated ovaries. We define this variable as the frequency of workers with ovaries in which the largest oocytes were >2 mm long (described as "'ready to lay' eggs" in *Amsalem, Orlova & Grozinger (2015)*), and in which the terminal oocyte was not resorbed (because resorption suggests that the worker will not lay a viable egg soon; *Duchateau & Velthuis, 1989*). This metric is also directly comparable to the one used in *Van Oystaeyen et al. (2014)*, which is one of the past studies that Amsalem et al. were following up. As evidence that this metric is biologically meaningful, we found that it was a good predictor of egg laying, and was a better predictor than other possible binary measures of ovary activation derived from Amsalem et al.'s dissection data (Table S3).

RESULTS OF OUR STATISTICAL ANALYSIS

Our reanalysis found evidence that one or more hydrocarbons significantly inhibited worker reproduction, for some but not all of the ovary metrics (Fig. 1; Tables S4–S8). Additionally, our reanalysis does not support claims in *Amsalem, Orlova & Grozinger (2015)* that the experiment had "a power of 99% and 96% for effect size of 0.2 and 0.5, respectively" (sic) to detect treatment effects. For example, Fig. 1 shows that the 95% confidence intervals are large, such that many of the non-significant results are consistent with large treatment effects. We have archived Amsalem et al.'s raw data, and the R scripts for our new analyses, alongside this article.

Following requests from reviewers, we analysed each dataset with multiple different modelling approaches, in order to verify that our results are robust to the choice of model

used. For example, we investigated the egg count data in Fig. 1A with Poisson generalised linear models (GLM; colony was modelled as a fixed effect), Poisson generalised linear mixed models (GLMM; colony modelled as a random effect), and a Bayesian Poisson generalised mixed model (see Tables S4–S8 and attached R script for full details). In every case, the qualitative conclusions were highly concordant, suggesting the differences with the conclusions in *Amsalem, Orlova & Grozinger (2015)* are not particular to some cherry-picked model. For brevity, Fig. 1 shows the results of our preferred analyses only (i.e., frequentist GLMMs and a mixed effects survival analysis) and omits the one response variable (oocyte resorption) for which there was no qualitative difference between our results and those in *Amsalem, Orlova & Grozinger (2015)* (namely, that all three queen CHCs induced significantly higher oocyte resorption).

Amount and timing of worker egg laying

Egg number and the latency to egg laying are arguably the most direct measures of worker reproduction. Amsalem et al. reported no effect of hydrocarbons on either variable using ANOVA and post-hoc testing. Our reanalysis suggested that the high dose of C_{25} caused worker groups to lay half as many eggs, and to take 55% longer to begin egg laying, relative to the control, in experienced workers (Poisson GLMM and mixed effects survival analysis—results agreed with GLM and Bayesian GLMM models; Figs. 1A–1B; Tables S4–S5). In addition, the C_{27}-high treatment delayed the onset of egg laying in experienced workers (Fig. 1B). All of these results remained significant after controlling for multiple testing (*Benjamini & Hochberg, 1995*).

Frequency of workers with fully activated ovaries

Among naïve workers, all three of the high-dose hydrocarbon treatments significantly reduced the proportion of workers with activated ovaries (binomial GLMM plus alternative analyses; Fig. 1C; Table S6). The effect size of hydrocarbons on the frequency of workers with active ovaries was of similar magnitude to that observed in previous comparable experiments in bumblebees, wasps and ants (but not honeybees) using analogous response variables (Table S2). All of these results remained significant after controlling the false discovery rate (*Benjamini & Hochberg, 1995*).

We speculate that the 'experienced' workers responded to the queen hydrocarbons with a reduction in egg laying, while the 'naïve' workers responded with a reduction in ovary activation only towards the end of the experiment, because of confounding differences in the age of these workers. The naïve workers all had undeveloped ovaries at the start of the experiment (since these workers were <24 h old), while the experienced workers were presumably in a variety of stages of reproductive development, priming them to begin egg laying sooner (as was observed: Fig. 1 and *Amsalem, Orlova & Grozinger, 2015*), and making the effects of pheromones on egg number more pronounced in the 'experienced' treatment (since egg laying occurred over more days in the experienced treatment than in the naïve treatment).

Mean size of terminal oocytes in workers' ovaries

Hydrocarbon treatment did not significantly affect the size of the terminal oocytes. Workers receiving the 'High' dose of the *B. terrestris* queen pheromone C_{25} had non-significantly smaller terminal oocytes than the control ($p = 0.077$ in a planned contrast, i.e., uncorrected for multiple testing; Figs. 1D; Table S7).

Frequency of oocyte resorption

We also replicated Amsalem et al.'s finding that all three hydrocarbons induced significantly more oocyte resorption than the hexane control (binomial GLMM; Table S8; not shown in Fig. 1 since the results are the same as in *Amsalem, Orlova & Grozinger (2015)*). Although this finding is consistent with the three CHCs somehow affecting ovaries, the biological significance is harder to interpret, because oocyte resorption in *B. impatiens* had a weak *positive* relationship with egg-laying (*Amsalem, Orlova & Grozinger, 2015*). This result contrasts with a previous study of a different *Bombus* species, in which oocyte resorption was more common in workers living in queenright rather than queenless colonies (Table S2D in *Van Oystaeyen et al. (2014)*), suggesting that queenless workers begin to reproduce because they cease resorbing their oocytes.

Effect of body size on measures of worker reproduction

In addition to the large effect of colony identity seen in most of the analyses (Table S3), worker body size was confirmed to have a significant effect on worker egg-laying and ovary development in nearly all analyses (Tables S4–S8). Specifically, cages with larger workers produced more eggs ($p = 0.001$) and laid earlier ($p = 0.0003$), and larger workers had bigger oocytes ($p < 0.0001$) and were more likely to display oocyte resorption ($p = 0.03$).

These results underscore the importance of controlling for body size, either experimentally or statistically, when studying reproduction in size polymorphic insects such as bumblebees. Out of curiosity, we re-analysed the data from our previous bumblebee queen pheromone experiment (*Van Oystaeyen et al., 2014*) after including body size as a covariate, in order to check the conclusions of *Van Oystaeyen et al. (2014)* were robust. The treatment effect of C_{25} on oocyte resorption remained, and we found that larger workers were less likely to have resorbed oocytes (Table S9).

CONCLUSIONS AND RECOMMENDATIONS

To conclude, we suggest that Amsalem et al.'s claim—that their experiment definitively demonstrates that three fertility-associated hydrocarbons do not reduce worker reproduction in *B. impatiens*—does not follow from their data. The experiment has an unbalanced and comparatively low sample size, which is problematic because the study's conclusion rests upon its failure to reject the null hypothesis. We also highlighted a number of methodological problems, such as confounding effects caused by non-random assignment of individuals to treatments, which complicate interpretation of the data.

Our reanalysis found evidence that C_{25}, the same queen pheromone identified in *B. terrestris* (*Van Oystaeyen et al., 2014*; *Holman, 2014*), substantially reduced the number of eggs laid, delayed the onset of laying, and reduced the frequency of workers with

activated ovaries in *B. impatiens*. We also found limited evidence that the other two fertility-associated hydrocarbons (C_{23} and C_{27}) might perform a similar function. The results are patchy (Fig. 1), and our reanalysis is not decisive because of the issues with the data. Nevertheless, the reanalysis makes it clear that the new study does not comprehensively reject the hypothesis that queen-like CHCs are not involved in regulating reproduction in *B. impatiens* workers, as claimed.

We suggest the following modifications to future experiments to help ensure reliable results. Firstly, an appropriate sample size is needed to ensure adequate statistical power, particularly when the effect sizes are expected to be moderate (see effect size estimates from past queen pheromone experiments in Table S2). Although we applaud the effort in *Amsalem, Orlova & Grozinger (2015)* to examine multiple chemicals, doses, and categories of workers, the workload needed to maintain an adequate sample size becomes prohibitive very quickly, and so it may be better to design experiments with good replication but fewer treatments. Secondly, one should start with a common pool of individuals and then randomly allocate them to treatments, rather than allocating different pools (e.g., young and old workers, or big and small workers) to different treatments, producing confounding effects. This can be done by splitting colonies randomly and equally between pheromone treatments (as in e.g., *Van Oystaeyen et al., 2014*; *Holman et al., 2010*; *Holman, Lanfear & D'Ettorre, 2013*; *Holman, Hanley & Millar, 2016*; *De Narbonne et al., 2016*; *Oi et al., 2016*), or randomly assigning whole colonies to different treatments (e.g., *Van Oystaeyen et al., 2014*; *Holman, 2014*). It is also important to run the different experimental treatments in parallel, rather than running one treatment and then another, such that environmental factors or cohort effects could confound the results (it is unclear whether this was done in *Amsalem, Orlova & Grozinger (2015)*, but the differences in sample size and worker colony origin imply that it was not). Thirdly, we acknowledge that it can be difficult to select the correct dose of pheromone in this type of study, since we can think of no foolproof way to accurately measure the dose to which workers are exposed in natural colonies. The ideal experiment may be to compare worker responses to natural queens, queens whose pheromone was somehow selectively removed) e.g., through using genetic manipulation), and appropriate control queens. *Korb et al. (2009)* performed such an experiment, in which they used RNAi to knock out a gene putatively involved in chemical communication in queen termites, and observed an increase in a worker behavior associated with queenlessness. The challenge for such experiments is to ensure that the only change in the queen is the removal of her pheromone. Alternatively, one could test multiple doses of pheromone that span the conceivable range of concentrations that workers might experience. Finally, one could test whether queen pheromones are learned by collecting naïve workers, giving them a 'training phase' with either no queen, a queen of low fecundity, or a queen of high fecundity, and then later measuring their physiological or behavioral responses to queen pheromone.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

Annette van Oystaeyen is an employee of Biobest Belgium, Westerlo, Belgium.

Author Contributions

- Luke Holman, Jelle S. van Zweden and Tom Wenseleers conceived and designed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Ricardo C. Oliveira and Annette van Oystaeyen conceived and designed the experiments, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:

The raw data has been supplied as a Supplemental Dataset.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.3332#supplemental-information.

REFERENCES

Amsalem E, Orlova M, Grozinger CM. 2015. A conserved class of queen pheromones? Re-evaluating the evidence in bumblebees (*Bombus impatiens*). *Proceedings of the Royal Society B: Biological Sciences* **282**:20151800 DOI 10.1098/rspb.2015.1800.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**:289–300.

De Narbonne MM, Van Zweden JS, Bello JE, Wenseleers T, Millar JG, D'Ettorre P. 2016. Biological activity of the enantiomers of 3-methylhentriacontane, a queen pheromone of the ant *Lasius niger*. *Journal of Experimental Biology* **219**(**Pt 11**):1632–1638 DOI 10.1242/jeb.136069.

D'Ettorre P, Heinze J, Schulz C, Francke W, Ayasse M. 2004. Does she smell like a queen? Chemoreception of a cuticular hydrocarbon signal in the ant *Pachycondyla inversa*. *Journal of Experimental Biology* **207**:1085–1091 DOI 10.1242/jeb.00865.

Dietemann V, Peeters C, Liebig J, Thivet V, Hölldobler B. 2003. Cuticular hydrocarbons mediate discrimination of reproductives and nonreproductives in the ant *Myrmecia gulosa*. *Proceedings of the National Academy of Sciences of the United States of America* **100**:10341–10346 DOI 10.1073/pnas.1834281100.

Duchateau MJ, Velthuis HHW. 1989. Ovarian development and egg laying in workers of *Bombus terrestris*. *Entomologia Experimentalis et Applicata* **51**:199–213 DOI 10.1111/j.1570-7458.1989.tb01231.x.

Holman L. 2014. Bumblebee size polymorphism and worker response to queen pheromone. *PeerJ* **2**:e604 DOI 10.7717/peerj.604.

Holman L, Dreier S, D'Ettorre P. 2010. Selfish strategies and honest signalling: reproductive conflicts in ant queen associations. *Proceedings of the Royal Society B: Biological Sciences* **277**:2007–2015 DOI 10.1098/rspb.2009.2311.

Holman L, Hanley B, Millar JG. 2016. Highly specific responses to queen pheromone in three *Lasius* ant species. *Behavioral Ecology and Sociobiology* **70**:387–392 DOI 10.1007/s00265-016-2058-6.

Holman L, Jørgensen CG, Nielsen J, D'Ettorre P. 2010. Identification of an ant queen pheromone regulating worker sterility. *Proceedings of the Royal Society B: Biological Sciences* **277**:3793–3800 DOI 10.1098/rspb.2010.0984.

Holman L, Lanfear R, D'Ettorre P. 2013. The evolution of queen pheromones in the ant genus *Lasius*. *Journal of Evolutionary Biology* **17**:1549–1558 DOI 10.1111/jeb.12162.

Korb J, Weil T, Hoffmann K, Foster KR, Rehli M. 2009. A gene necessary for reproductive suppression in termites. *Science* **324**:758–758 DOI 10.1126/science.1170660.

Liebig J, Peeters C, Oldham NJ, Markstadter C, Hölldobler B. 2000. Are variations in cuticular hydrocarbons of queens and workers a reliable signal of fertility in the ant Harpegnathos saltator? *Proceedings of the National Academy of Sciences of the United States of America* **97**:4124–4131 DOI 10.1073/pnas.97.8.4124.

Monnin T, Malosse C, Peeters C. 1998. Solid-phase microextraction and cuticular hydrocarbon differences related to reproductive activity in queenless ant *Dinoponera quadriceps*. *Journal of Chemical Ecology* **24**:473–490 DOI 10.1023/A:1022360718870.

Oi CA, Millar JG, Van Zweden JS, Wenseleers T. 2016. Conservation of queen pheromones across two species of vespine wasps. *Journal of Chemical Ecology* **42**:1175–1180 DOI 10.1007/s10886-016-0777-9.

Smith AA, Hölldobler B, Liebig J. 2009. Cuticular hydrocarbons reliably identify cheaters and allow enforcement of altruism in a social insect. *Current Biology* **19**:78–81 DOI 10.1016/j.cub.2008.11.059.

Smith AA, Millar JG, Suarez AV. 2015. A social insect fertility signal is dependent on chemical context. *Biology Letters* **11**:20140947 DOI 10.1098/rsbl.2014.0947.

Van Oystaeyen A, Oliveira RC, Holman L, Van Zweden JS, Romero C, Oi CA, D'Ettorre P, Khalesi M, Billen J, Wäckers F, Millar JG, Wenseleers T. 2014. Conserved class of queen pheromones stops social insect workers from reproducing. *Science* **343**:287–290 DOI 10.1126/science.1244899.

Supplementary Methods

To estimate the amount of cuticular compounds of *Bombus impatiens*, we collected queens and workers from ten colonies obtained from BioBest Canada Ltd. (Leamington, Ontario, Canada). CHCs were extracted by immersing individual bumble bees in HPLC-grade pentane (Sigma-Aldrich, Belgium), 5 ml in case of queens or 2 ml in case of workers, for 5 min with vortexing at 600 rpm for 10 sec at the start and the end. The pentane was then evaporated, and all extracts were reconstituted in 1 ml HPLC-grade hexane (Sigma-Aldrich, Belgium), and transferred to 2 ml vials. We then injected 0.5 µl of extract into a Thermo Scientific TRACE 1300 gas-chromatograph coupled with a Thermo Scientific ISQ mass-spectrometer. Injection was done in split mode (10:1), using an injection temperature at 320 °C, MXT-5 stainless steel capillary column (30 m × 0.25 mm × 0.25 µm), helium as the carrier gas at 0.9 ml/min, a MS transfer-line temperature of 300 °C, and an ion source temperature of 300 °C. The oven temperature was held at 40 °C for 2 min, then increased to 120 °C at 20 °C/min, then to 200 °C at 10 °C/min, then to 250 °C at 7 °C/min, and then to 350 °C at 5 °C/min, with a final hold of 15 min at 350 °C. AMDIS 2.71 in combination with NIST MS Search 2.2, NIST 2014 library search, and manual interpretation of mass spectra (taking into account expected fragmentation patterns and retention indices) were used to identify the different compounds. Peak areas of 29 hydrocarbons were integrated using a custom script (available from the authors upon request) in R v3.2.4. Retention indices and absolute quantities of our compounds were calculated using cubic spline interpolation based on the elution times and peak areas of an external n-C7 to n-C40 linear alkane ladder standard (49452-U, Supelco) that was run on the same program with concentrations of 0.1, 0.01 and 0.001 µg/µl.

The results of this analysis can be found in Supplementary Table S11.

Legends for supplementary tables

Table S1: ANOVA analyses demonstrating that worker body size significantly differed across colonies and across the hydrocarbon and naivety treatments, and therefore should be included as a covariate in subsequent analyses to minimise confounding effects.

Table S2: Overview of the methods and results of experimental studies investigating the effects of synthetic putative queen pheromones on worker ovaries or reproduction. In all cases, the original data was requested (allowing us to recalculate effect size in a standardized way – from a GLMM with colony as a random factor), but if these were not available the frequencies of workers with developed ovaries were estimated from the published papers. The papers were found using an exhaustive search on Google Scholar.

Table S3: Analysis to find the ovary metrics that best predict the amount of worker egg-laying. Table S3a gives the results of a quasi-Poisson GLM with colony treated as a fixed factor, whereas Table S3b gives those of a Poisson generalized linear mixed model, where cage ID and colony ID are treated as random factors in the model (the cage ID random effect was included to correct for overdispersion). In both cases, the proportion of workers with developed and non-resorbed, developed and resorbed, and non-developed and resorbed ovaries were included as covariates, whilst correcting for the total number of days over which eggs were laid in each cage. The fourth category (i.e. non-developed and non-resorbed) was left out to prevent the model from becoming rank deficient, since the proportion of workers in each category sums to one. Workers with developed ovaries were scored as those with a mean terminal oocyte size of >2 mm. The proportion of workers with developed and non-resorbed oocytes was the strongest predictor of actual egg-laying.

Table S4: Results of three different models of the total number of eggs produced over 10 days in cages containing 3 workers. Tables S4a-c give the results from a Poisson GLMM with colony treated as a random factor (allowing all the data to be analysed in a single model), the experienced/naïve treatment as a fixed factor and mean-centered group mean body size as a covariate. Tables S4d and S4e describe the results of a pair of quasi-Poisson GLMs, which separately analyse the data for the naïve and experienced workers (splitting the dataset like this was necessary because colony was confounded with the experience treatment). Colony and treatment were entered as fixed effects, and mean-centered group mean body size was included as a covariate. Finally, Table S4f gives the results of a Poisson GLMM conducted in a Bayesian framework using the *brms* package for R. The results of the three modeling approaches are all broadly concordant: all three models suggest that C_{25} significantly reduced the number of eggs laid relative to the hexane control. All models also showed that cages that contained larger workers laid significantly more eggs.

Table S5: Tables S5a-c describe the results of an analysis of the latency for workers to begin laying eggs based on a Weibull frailty model with treatment, worker type and treatment × worker type as fixed effects, mean-centered group

mean body size of the workers as a continuous covariate, and colony ID as a Gaussian frailty term. The tables show the summary coefficient table, least square means for treatment and worker type, and the response ratio for each hydrocarbon treatment relative to the hexane control (based on posthoc tests with FDR correction of *p* values for multiple testing). Tables S5d-S5e show the results of a similar model but with colony entered as a fixed factor instead of a frailty term. In this case, however, separate models were run for naïve and experienced workers because colony and experienced were confounded and the model could not handle this elsewise. Table S5f shows the results of a pair of Bayesian proportional hazards models (one for naïve workers, one for experienced workers), fitting the effects of hydrocarbon treatment and mean body size, as well as a Gaussian frailty term (fit using the using the *brms* package for R). Both the frequentist and Bayesian models found that C_{25} significantly increased the time taken by workers to begin laying eggs, and that cages with larger workers started laying eggs earlier.

Table S6: Results of three different models of the frequency of workers with active ovaries, which as in [4] we defined as containing ready-to-lay eggs in their ovaries (i.e. containing fully mature oocytes with a terminal oocyte size >2 mm with no signs of oocyte resorption). Tables S6a-c show the results of a binomial GLMM with treatment, worker type and treatment × worker type as fixed effects, mean-centered worker body size as a continuous covariate, and cage ID and colony ID as random intercepts. Results show the summary coefficient table, least square means for treatment and worker type, and the response ratio for each hydrocarbon treatment relative to the hexane control (based on posthoc tests with FDR correction of *p* values for multiple testing). Tables S6d-e show the results of a pair of binomial generalized linear models (one for experienced workers, one for naïve workers; as before, the data needed to be split because colony and treatment were confounded) in which treatment and colony were included as fixed factors and mean-centered body size as a covariate. Table S6f shows a binomial GLMM conducted in a Bayesian framework using the *brms* package for R. Finally, Tables S6g-h and S6i show the results of a fixed-effect and of a random-effect ordinal logistic regression in which the 4 possible ovary development scores were analyzed on an ordinal scale (in increasing order of ovary development: 1=non-developed and non-resorbed, 2=non-developed and resorbed, 3=developed and resorbed and 4=developed and non-resorbed). The results of all five modeling approaches are all broadly concordant, and both the Bayesian and frequentist binomial mixed models as well as the ordinal models suggest that the three "high" treatments reduced the frequency of naïve workers with active ovaries.

Table S7: Analysis of the mean size of the terminal oocytes in the workers' three largest ovarioles. The distribution of these data is truncated (since an egg cannot be smaller than zero, nor much larger than an average-sized mature egg) and exhibits bimodality, so we rescaled the oocyte size data to lie between 0 and 1 and then analyzed them using a beta-distributed generalized additive mixed model with logit link function. We selected the most parsimonious model using the Aikaike Information Criterion, which contained treatment, worker type (experienced or naïve) and mean-centered worker body size (but not the

treatment-type interaction); we also included cage ID and colony ID as random effects in all models compared. Results show the Type III ANOVA table, the summary coefficient table (which shows contrasts with the hexane control), and the least square means for treatment and worker type.

Table S8: Analysis of the proportion of workers displaying oocyte resorption. The model is a binomial GLMM with treatment, worker type and treatment × worker type as fixed effects, mean-centered worker body size as a continuous covariate, and cage ID and colony ID as random intercepts. Results show the summary coefficient table, least square means for treatment and worker type, and the response ratio for each hydrocarbon treatment relative to the hexane control (based on posthoc tests with FDR correction of *p* values for multiple testing). Analysing the data with Bayesian GLMM or GLM yielded qualitatively identical results. These results are not illustrated in Fig. 1, since our conclusions concur with Amsalem et al.'s for these data.

Table S9: Reanalysis of the queen pheromone bioassays of Van Oystaeyen et al. [4] on the proportion of workers displaying oocyte resorption or having active ovaries in *B. terrestris* using binomial GLMMs with explicit consideration of worker body size (which was not considered in [4] but which we have measured since). As above, active ovaries were defined as containing ready-to-lay eggs (i.e. with a mature terminal oocyte with no signs of oocyte resorption; ovary development scale IV of [18]). The models used are binomial GLMMs with treatment included as a fixed factor and mean-centered colony size and mean-centered worker body size (thorax width) as covariates, and with colony ID as a random intercept. Results show the summary coefficient table and the response ratio for each hydrocarbon treatment relative to the hexane control (based on posthoc tests with FDR correction of *p* values for multiple testing). Results confirm the previously reported effect that C_{25} causes a significant increase in oocyte resorption [4], which in this species has been shown to be linked to suppression of worker reproduction [4], and that larger workers are significantly more likely to have active ovaries and less likely to display oocyte resorption.

Table S10: Reanalysis of the queen pheromone bioassays of Holman [10] of the the number of visible oocytes present in the workers' ovaries and of the proportion of workers having developed ovaries (defined as containing more than 10 visible oocytes) in *B. terrestris* using negative binomial and binomial GLMMs and with explicit consideration of worker body size (here simply coded as small, medium or large). The models used are binomial or negative binomial GLMMs with treatment included as a fixed factor and mean-centered worker body size as a covariate, and with colony ID as a random intercept. Results show the summary coefficient table of each analysis (here no FDR correction of *p* values for multiple testing was required as there was only one treatment). Results confirm the previously reported effect that C_{25} causes a significant decrease in visible oocyte number [10] and a decrease in the probability for workers to have developed ovaries, and also show that larger workers had more visible developing oocytes in their ovaries and were more likely to have developed ovaries.

Table S11: The mass (in μg) of each hydrocarbon detected on the cuticles of 10 queens and 20 workers of *Bombus impatiens*.